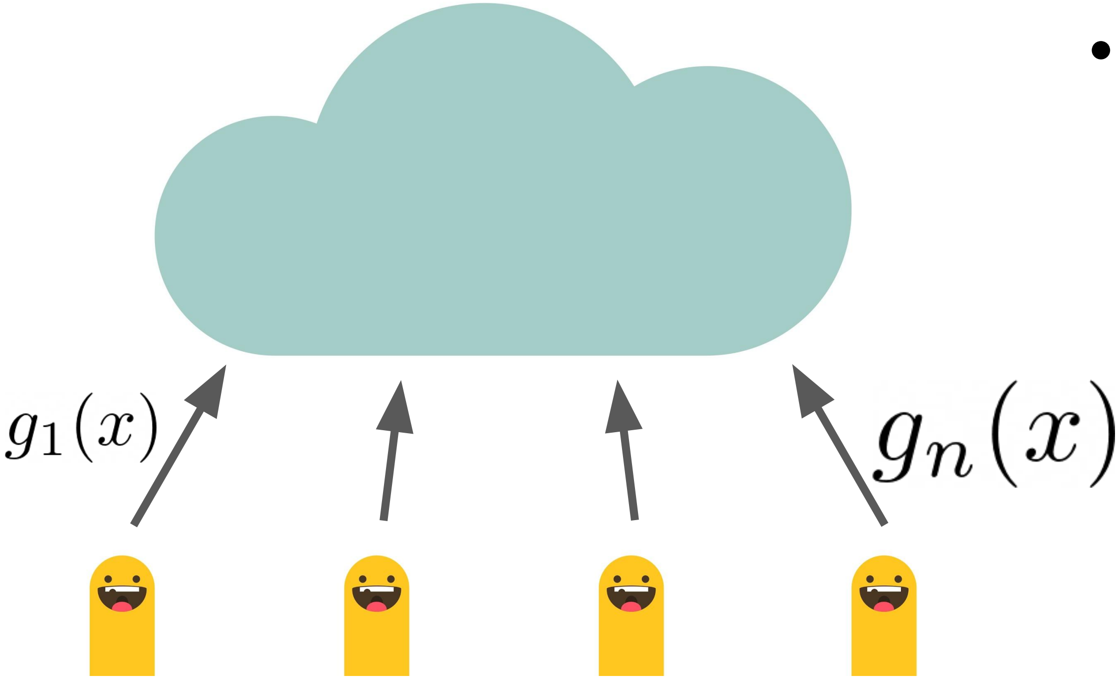
EPFL

Learning from history for Byzantine robust optimization

Sai Praneeth Karimireddy, Lie He, Martin Jaggi

Byzantine robust learning

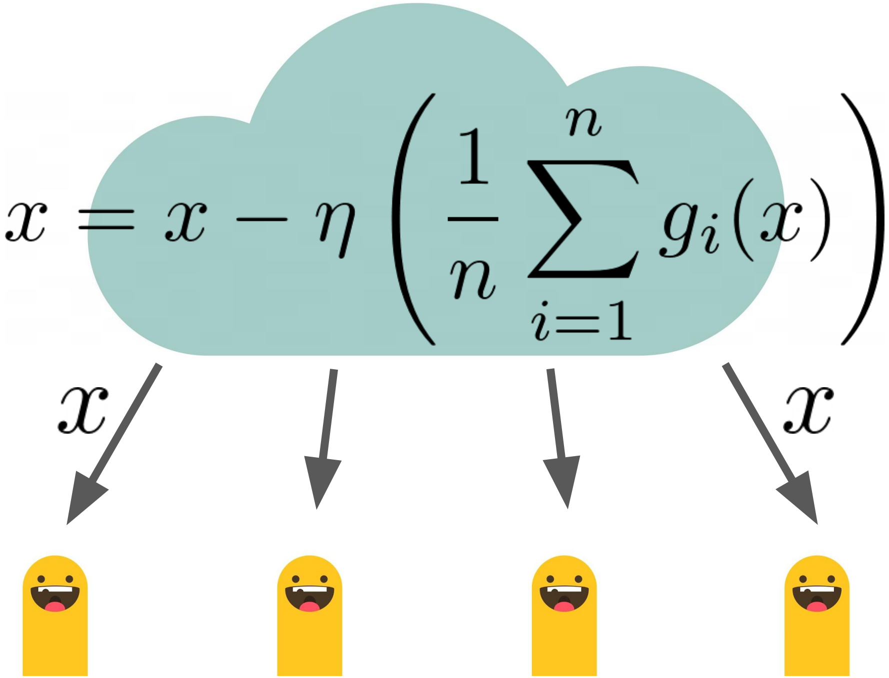
Federated learning



- Each worker i computes stochastic gradient at x and sends to server

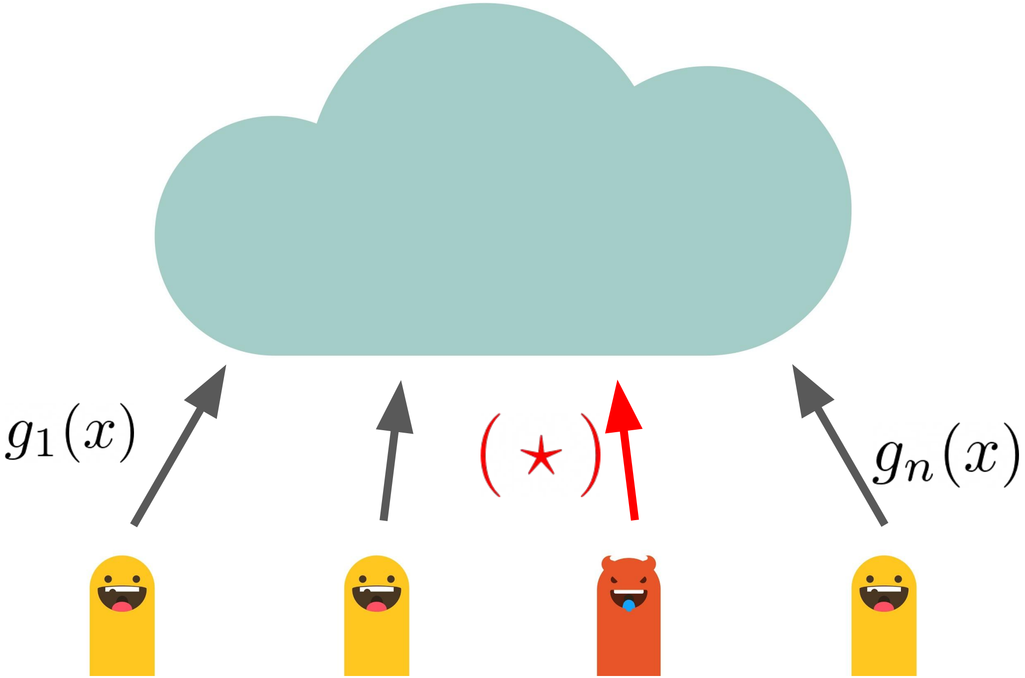
Federated learning

$$x = x - \eta \left(\frac{1}{n} \sum_{i=1}^{n} g_i(x) \right)$$



- Server accumulates gradients and computes new parameters

Byzantine robust learning



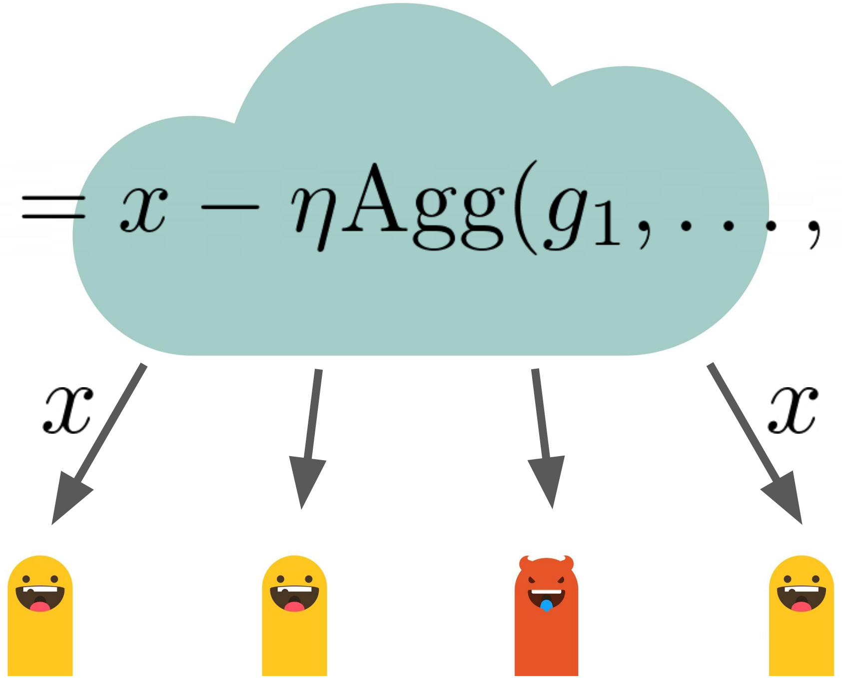
We protect against worst-case:

- Small fraction (δ) of workers may send arbitrary updates
- They are omniscient and can collude
- They want to derail convergence

Failure of current aggregators

Classic robust algorithms

$$x = x - \eta \mathrm{Agg}(g_1, \ldots, g_n)$$

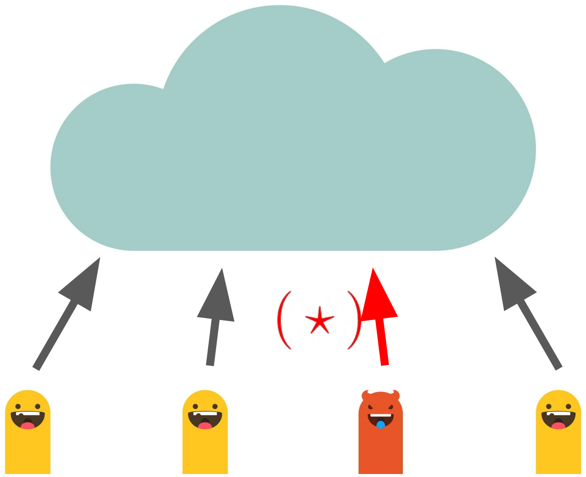


- Replace Avg with different aggregator

Examples:

- Coordinate-wise median [Yin et al. 2017]
- Krum [Blanchard et al. 2018]
- Geometric median / RFA [Pillutla et al. 2019]

Median based aggregators



Simplest is perhaps coordinate wise median [Yin et al. 2018]

*K*th coordinate is computed as:

$$[\mathrm{CM}(g_1, \ldots, g_n)]_k = \mathrm{median}([g_1]_k, \ldots, [g_n]_k)$$

Median based aggregators: theoretical failure

- Consider all correct outputs

  (+1, -1, +1, -1, +1, ..., -1)

- Correct Avg is 0

- Median outputs ±1

$$[\mathrm{CM}(g_1, \dots, g_n)]_k = \mathrm{median}([g_1]_k, \dots, [g_n]_k)$$

Median based aggregators: theoretical failure

- Consider all correct outputs

  (+1, -1, +1, -1, +1, …., -1)

$$[\mathrm{CM}(g_1, \ldots, g_n)]_k = \mathrm{median}([g_1]_k, \ldots, [g_n]_k)$$

- Correct Avg is 0

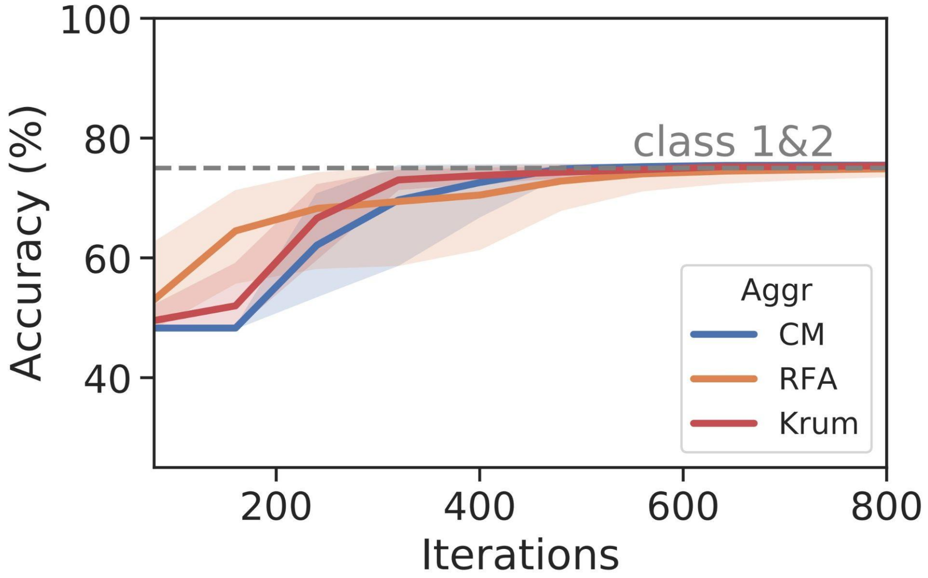
- Median outputs ±1

- CM, Krum, RFA all fail in more general settings
  (see paper for theory)

Median based aggregators: experimental failure

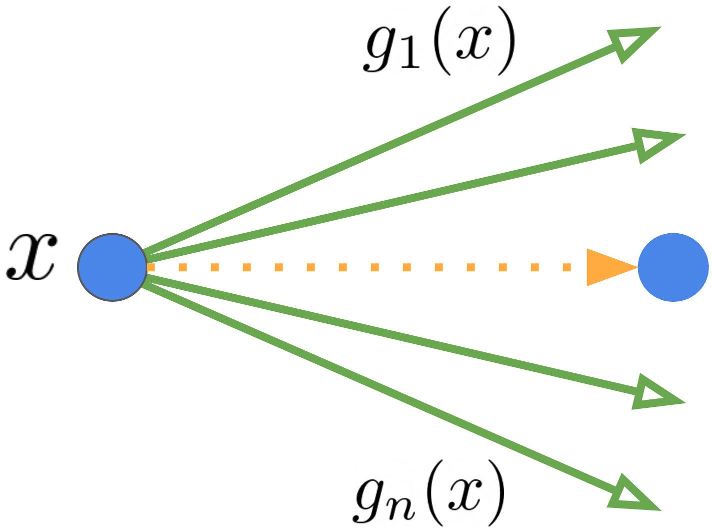
- We construct long-tailed MNIST dataset

- 75% accuracy corresponds to learning only class 1 & 2 and ignoring all others.



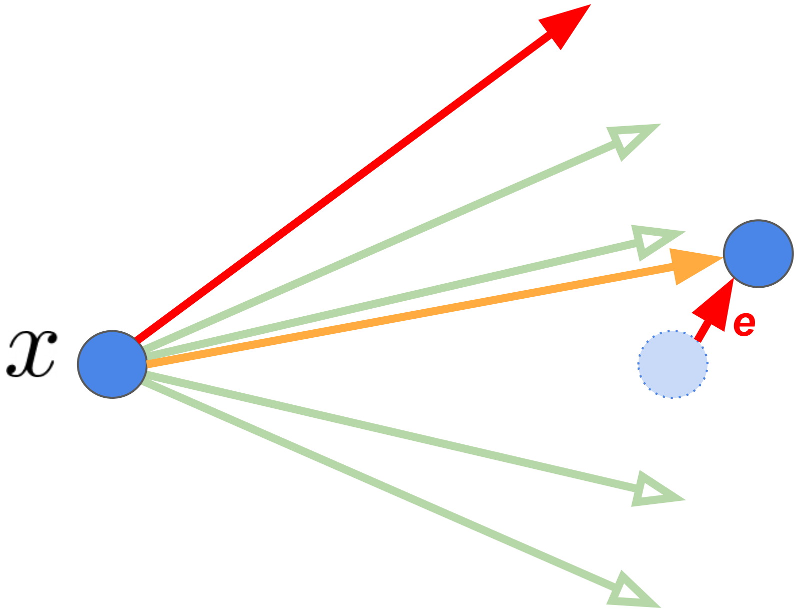
Necessity of history

Necessity of history: ideal update



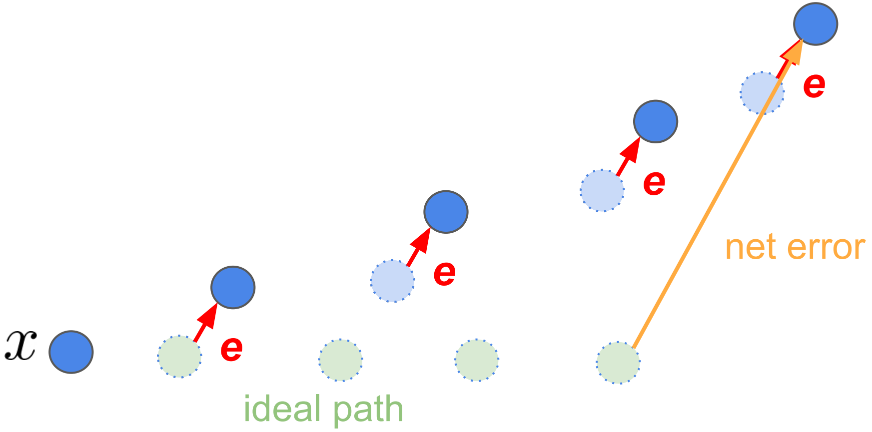
$$x = x - \eta \left(\frac{1}{n} \sum_{i=1}^{n} g_i(x) \right)$$

Necessity of history: approximate update



- Suppose, we **successfully** defend Byzantine attacks with smaller error ***e***

Necessity of history: approximate update



- Attacks can be ***coupled across time***
- Error ***e*** adds up over time
- Eventually, leads to large divergence

Necessity of history: theorem

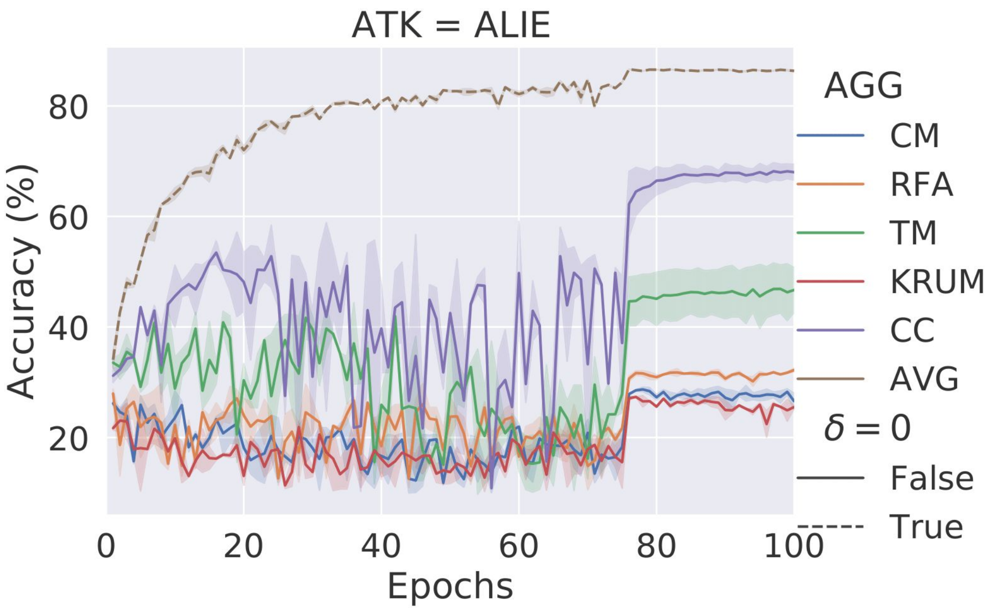
Impossible to avoid for any algorithm which is ‘memory-less’:

Theorem: For a μ-strongly convex function, the output ***x*** of any memory-less algorithm necessarily has an error:

$$\Omega\left(\frac{\delta\sigma^2}{\mu}\right)$$

Necessity of history: experiment

- “A little is enough” (ALIE) attacks on normal MNIST.
- Dotted line is ideal accuracy
- All aggregators (solid lines) have 20--60% drop in accuracy



Robust aggregator:
centered clipping

Robust aggregator: new definition

(δ_{max}, c)-robust aggregator:

For $\delta < \delta_{max}$, suppose **(1-δ)** fraction of inputs are good and satisfy

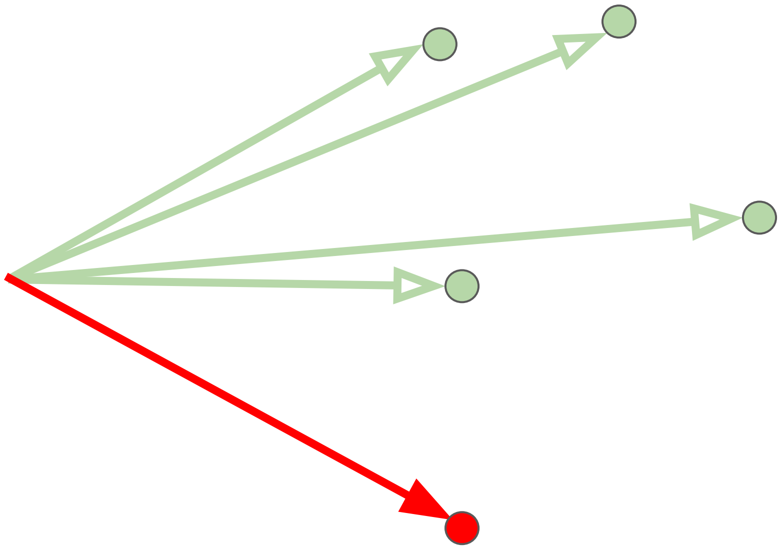
$$\mathbb{E}\|x_i - x_j\|^2 \leq \rho^2$$

Then, the output of the aggregator $\mathbf{x}^{out}$ **= Agg ($\mathbf{x}_1$,...., $\mathbf{x}_n$)** satisfies

$$\mathbb{E}\|x^{\text{out}} - \bar{x}\|^2 \leq c\delta\rho^2$$

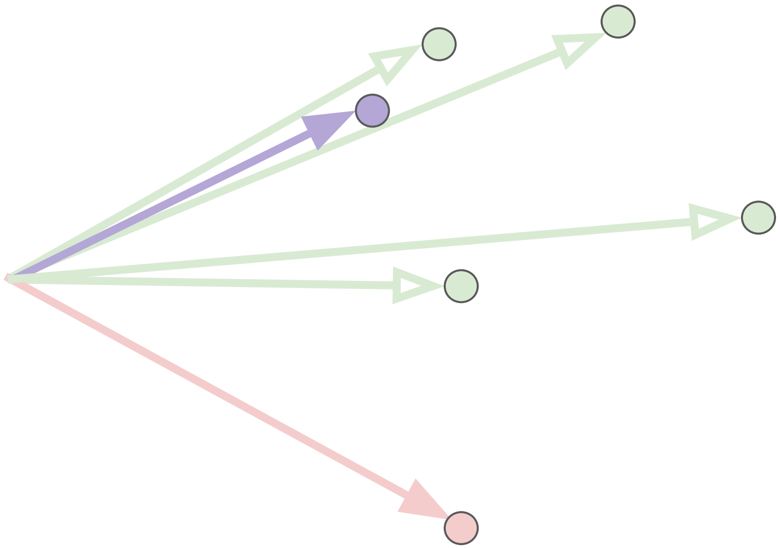
- If **δ**=0, then error is 0
  Median is not a robust aggregator.
- If ϱ=0, then error is 0
- Turns out this is best we can do (see paper)

Robust aggregator: centered clipping



Suppose we are give some inputs

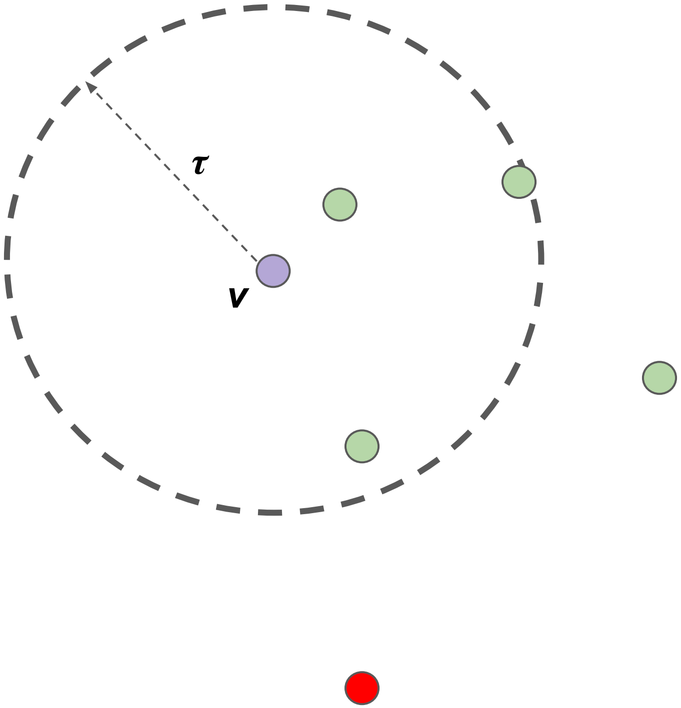
Robust aggregator: centered clipping



Suppose we are give some inputs

And a “guess” ***v***,

Robust aggregator: centered clipping

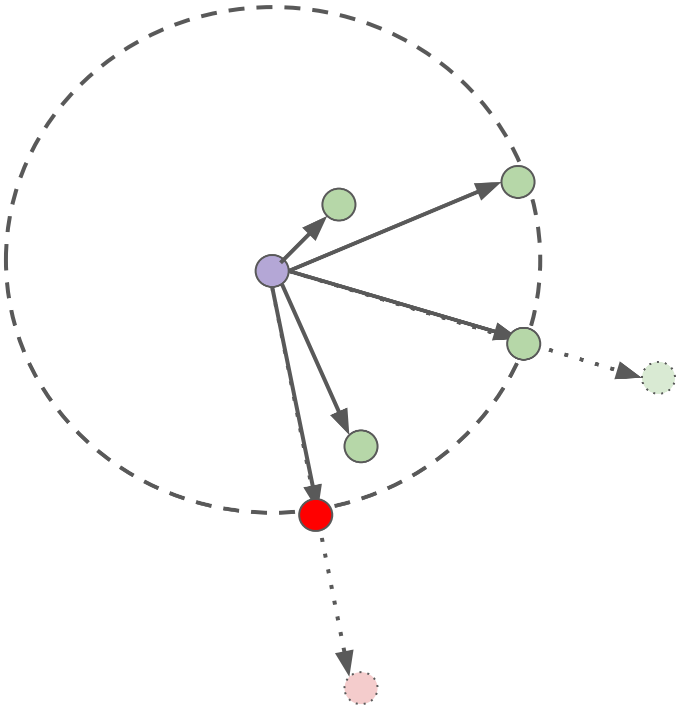


Suppose we are give some inputs

And a “guess” $\boldsymbol{v}$,

And clipping threshold $\boldsymbol{\tau}$.

Robust aggregator: centered clipping



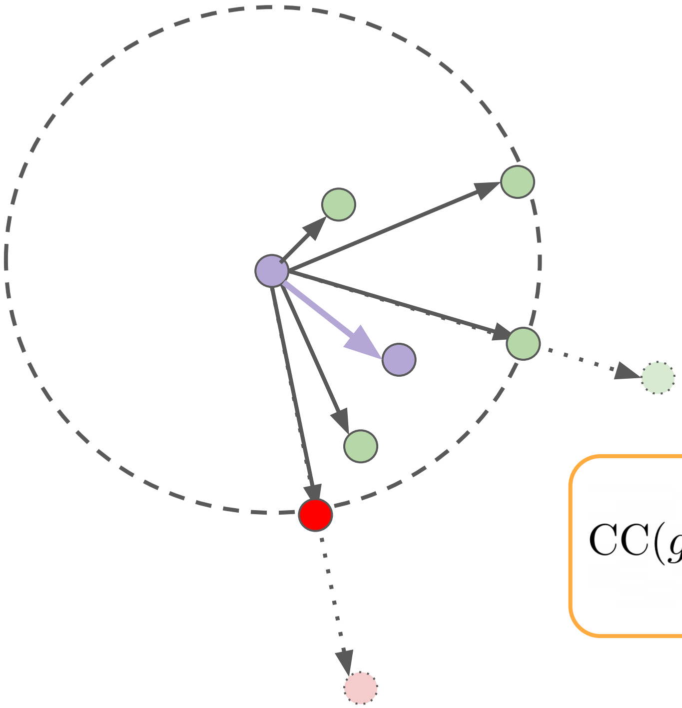
Suppose we are give some inputs

And a “guess” $\boldsymbol{v}$,

And clipping threshold $\boldsymbol{\tau}$.

Clip all values from guess to clipping threshold and average

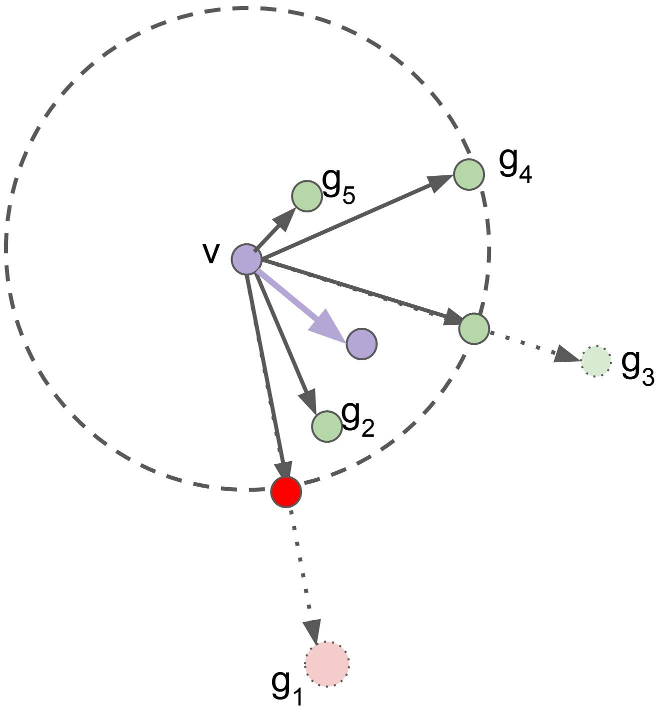
Robust aggregator: centered clipping



Clip all values from guess to clipping threshold and average

$$\mathrm{CC}(g_1, \ldots, g_n) = v + \frac{1}{n} \sum_{i=1}^{n} \mathrm{clip}_\tau (g_i - v)$$

Robust aggregator: theory

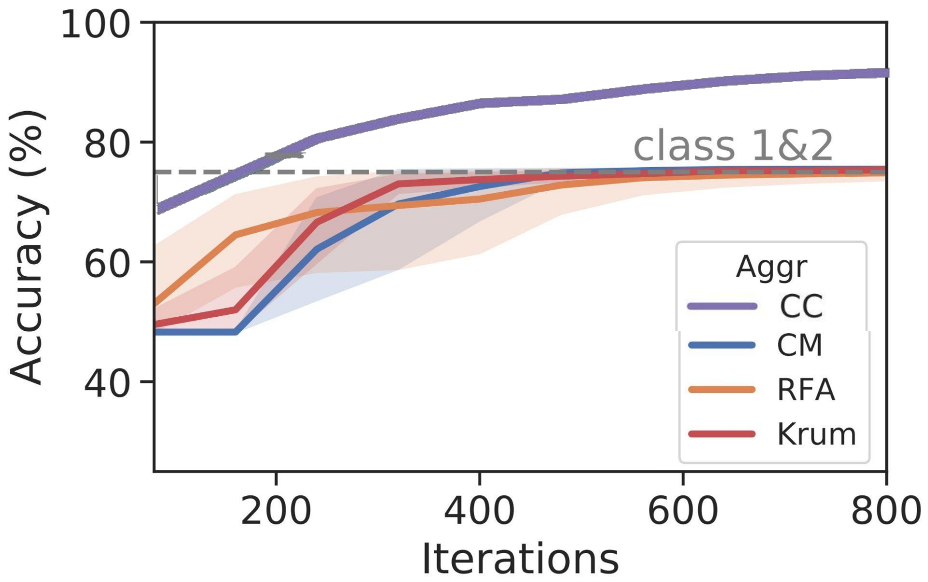


Theorem. Given a good starting point ***v***, centered clip is a **($\boldsymbol{\delta}_{max}$, c) robust aggregator** for $\boldsymbol{\delta}_{max}$ = 0.15 and c = O(1).

$$\mathrm{CC}(g_1, \ldots, g_n) = v + \frac{1}{n} \sum_{i=1}^{n} \mathrm{clip}_\tau(g_i - v)$$

Roust aggregator: experiment

- Long-tailed MNIST dataset
- Centered clip beats all other methods
- For guess ***v***, use aggregate output of previous round



Time coupled attacks:
worker momentum

Using history: momentum

- Simply use worker momentum

$$m_i = (1 - \beta)g_i + \beta m_i$$

- Effectively ***averages*** past gradients, reducing variance

- Aggregate worker momentums instead of gradients

$$x = x - \eta \text{Agg}(m_1, \ldots, m_n)$$

Using history: convergence theory

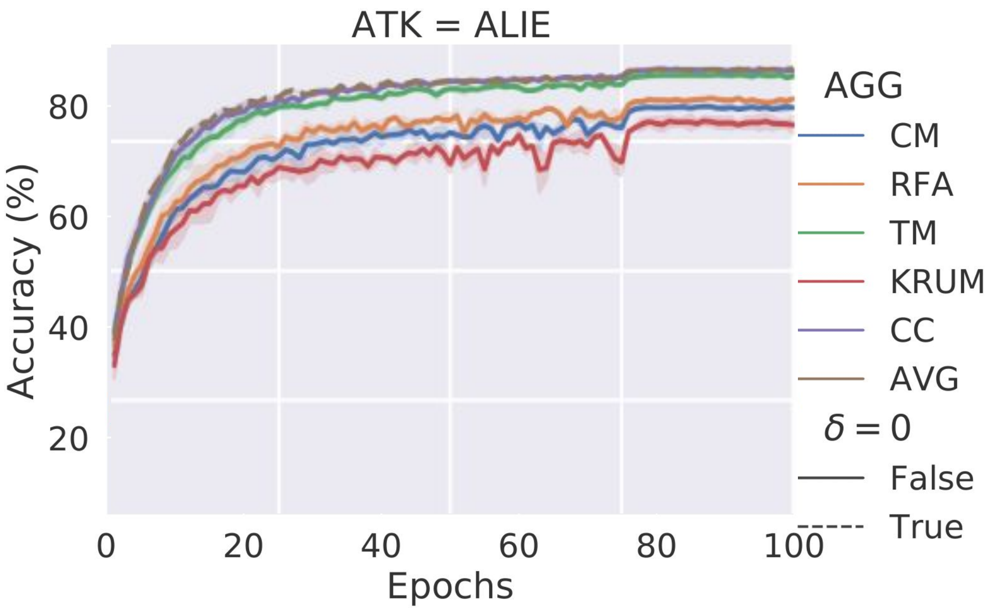
Theorem: Given any **($\boldsymbol{\delta}_{max}$, c) robust aggregator**, and a Byzantine robust problem with **$\boldsymbol{\delta}$**-fraction attackers and **$\boldsymbol{\sigma}^2$** variance, our algorithm outputs **$\mathbf{x}^{out}$** s.t.

$$\mathbb{E}\|\nabla f(x^{\text{out}})\|^2 \leq \mathcal{O}\left(\sqrt{\frac{\sigma^2}{T}\left(\delta + \frac{1}{n}\right)}\right)$$

Using history: experiment

- “A little is enough” (ALIE) attacks on normal MNIST with 0.99 momentum

- Centered Clip + momentum=0.99 matches ideal performance



Take-aways

1. Surprising failures can hide under assumptions

2. Need to use history for Byzantine robustness

3. Centered clipping with worker momentum provably and practically defends against Byzantine attacks